DOCUMENT RESUME

ED 451 634 EC 308 316

AUTHOR	Thurlow, Martha L.; McGrew, Kevin S.; Tindal, Gerald; Thompson, Sandra J.; Ysseldyke, James E.; Elliott, Judy L.
TITLE	Assessment Accommodations Research: Considerations for Design and Analysis. Technical Report 26.
INSTITUTION	National Center on Educational Outcomes, Minneapolis, MN.; Council of Chief State School Officers, Washington, DC.; National Association of State Directors of Special Education, Alexandria, VA.
SPONS AGENCY	Special Education Programs (ED/OSERS), Washington, DC.
PUB DATE	2000-12-00
NOTE	39p.
CONTRACT	H326G000001
AVAILABLE FROM	National Center on Educational Outcomes, University of Minnesota, 350 Elliott Hall, 75 East River Road, Minneapolis, MN 55455 ($15). Tel: 612-624-8561; Fax: 612-624-0879; Web site: http://www.coled.umn.edu/NCEO.
PUB TYPE	Reports - Research (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Disabilities; *Educational Assessment; Elementary Secondary Education; *Inclusive Schools; *Research Design; Research Methodology; *Research Needs; School Districts; Scores; State Programs; Test Validity
IDENTIFIERS	*Testing Accommodations (Disabilities)

ABSTRACT

This monograph reviews issues in research on the effects of using accommodations for students with disabilities who are included in educational assessments, particularly state and district assessments. Following an introductory overview, a section provides background on the need for good research on accommodations. The paper then defines commonly used terms and considers the purpose of assessment accommodations. Identification of issues in accommodations research considers comparability of scales administered under standard and nonstandard conditions, comparability of scores, and determination of cutoff scores. The paper explains three general analytic strategies in the context of accommodations research: item response theory, factor analysis, and criterion related analyses. Also considered are group research design considerations such as sampling and sample size. A major section explains four research designs in order from the most optimal to the least optimal. Other research designs considered include single subject, withdrawal-reversal, multiple baseline, multiple probe, changing criterion, and comparative designs. The paper concludes with four recommendations to researchers: (1) focus on the accommodation/s of most interest; (2) focus on those students who comprise the largest part of the population with disabilities; (3) if comparing groups, one group should be students without disabilities; and (4) collect other measures to help clarify findings. (Contains 30 references.) (DB)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Technical Report 26

ED 451 634

Assessment Accommodations Research: Considerations for Design and Analysis

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

NATIONAL CENTER ON EDUCATIONAL OUTCOMES

In collaboration with:

Council of Chief State School Officers (CCSSO)

National Association of State Directors of Special Education (NASDSE)

EC 308316

BEST COPY AVAILABLE

Technical Report 26

Assessment Accommodations Research: Considerations for Design and Analysis

Martha L. Thurlow • Kevin S. McGrew • Gerald Tindal •
Sandra J. Thompson • James E. Ysseldyke • Judy L. Elliott

December 2000

All rights reserved. Any or all portions of this document may be reproduced and distributed without prior permission, provided the source is cited as:

Thurlow, M. L., McGrew, K. S., Tindal, G., Thompson, S. J., Ysseldyke, J. E., & Elliott, J. L. (2000). *Assessment accommodations research: Considerations for design and analysis* (Technical Report 26). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

NATIONAL CENTER ON EDUCATIONAL OUTCOMES

The Center is supported through a Cooperative Agreement (#H326G000001) with the Research to Practice Division, Office of Special Education Programs, U.S. Department of Education. The Center is affiliated with the Institute on Community Integration at the College of Education and Human Development, University of Minnesota. Opinions expressed herein do not necessarily reflect those of the U.S. Department of Education or Offices within it.

NCEO Core Staff

John S. Bielinski
Jane L. Krentz
Michael L. Moore
Rachel F. Quenemoen
Dorene L. Scott
Sandra J. Thompson
James E. Ysseldyke

Martha L. Thurlow, Director

Additional copies of this document may be ordered for $15.00 from:

National Center on Educational Outcomes
University of Minnesota • 350 Elliott Hall
75 East River Road • Minneapolis, MN 55455
Phone 612/624-8561 • Fax 612/624-0879
http://www.coled.umn.edu/NCEO

The University of Minnesota is committed to the policy that all persons shall have equal access to its programs, facilities, and employment without regard to race, color, creed, religion, national origin, sex, age, marital status, disability, public assistance status, veteran status, or sexual orientation.

This document is available in alternative formats upon request.

Acknowledgments

The completion of this report by NCEO staff and others is a direct result of the encouragement and assistance provided by the Research Study Group of the Assessing Special Education Students (ASES) State Collaborative on Assessment and Student Standards (SCASS), one of many collaboratives supported by the Council of Chief State School Officers.

Representing the ASES Research Study Group, chairperson Patricia Almond writes:

The Research Study Group appreciates the membership of NCEO, and the many contributions that it makes to the group s concerns and deliberations. The Study Group, its states, and the ASES SCASS at large are eager to support NCEO and keep the momentum of its work going in any way possible. We hope that this is reflected in our support and contributions to this report.

Pat Almond

Executive Summary

The need for good research on the effects of assessment accommodations has exploded during the past five years. With the enactment of IDEA '97 has come an urgency to know whether the use of certain accommodations threatens test validity or score comparability. Similarly, there is a need to know whether specific accommodations are useful for individual students. Rigorous research designs are needed to ensure that accommodation research findings are useful to states and districts.

We wrote this report for researchers and for those who use research findings, as well as to potentially benefit IEP teams. It provides an overview of several group and single subject research designs. In addition, general analytic strategies are identified and explained in the context of accommodations research. These are item response theory (IRT), factor analysis, and criteria-related analysis.

We conclude with the following recommendations for accommodations research:

1. Focus on accommodations of most interest.
2. Focus on students who comprise a large part of the population needing accommodations, but do not use disability category as a proxy for the need for a specific accommodation.
3. If a comparison group design is used, at least one comparison group must be made up of students with no disabilities.
4. Collect other measures to help clarify findings.

It is an expectation that there will continue to be a need for accommodations research for some time to come. Hopefully, with this guide and others that follow, the research that is conducted will be both useful and informative.

Table of Contents

Overview 1
Background on the Need for Good Research on Accommodations 2
Definition of Terms 4
Purpose of Assessment Accommodations 5
Issues in Accommodations Research: Critical Research Questions 6
General Analytic Strategies 8
- Item Response Theory (IRT) 8
- Factor Analysis 10
- Criterion-Related Analyses 12

Group Research Design Considerations 13
- Sampling 13
- Size 14

Group Research Designs 14
- Design 1 14
- Design 2 15
- Design 3 17
- Design 4 17

Single Subject Research Designs 18
- Withdrawal-Reversal Design 20
- Multiple Baseline Designs 21
- Multiple Probe Designs 23
- Changing Criterion Designs 23
- Comparative Designs 25

Conclusions and Recommendations 26
References 29

Overview

This paper was written to address the tremendous need for good research on the effects of using accommodations during assessments, particularly state and district assessments. These assessments are used increasingly for high stakes purposes, with significant consequences for the student or for schools, administrators, and their staff (Heubert & Hauser, 1999). Thus, it is critical that research-based recommendations about whether an accommodation is appropriate to use be based on good research designs.

This paper was written for state directors of assessment, test developers, and researchers interested in conducting good accommodations research or being informed users of the results of accommodations research. It is our hope that through this paper, researchers will be aware of needed considerations for constructing good accommodations research designs, and that reviewers of completed research will know what to look for in the design of the research that has been completed.

All of this is more critical than ever now because the 1997 reauthorization of the Individuals with Disabilities Education Act (IDEA) requires that states and districts include students with disabilities in their assessments, with appropriate accommodations when necessary. Similar requirements emerge from the Title 1 provisions of the Improving America's Schools Act (known formerly as the Elementary and Secondary Education Act).

It is defining what comprise "appropriate" accommodations, both in terms of identifying what accommodations are needed by individual students, and in terms of the effect of accommodations on what is measured, that is at the heart of many concerns about the validity of accommodated assessments. It is because of this concern that there is a need for good research on accommodations. Defining "good" by considering the pros and cons of various research designs is at the heart of the purpose for developing this report.

While this paper was developed for research on accommodations for students with disabilities, much of it will apply as well to students with limited English proficiency (LEP). Despite this applicability, however, those conducting research on accommodations for LEP students should be aware that there are additional complicating factors that will confound the use of some of the research designs presented here. Primary among these confounding factors is an array of language issues that require additional considerations.

Background on the Need for Good Research on Accommodations

The participation of students with disabilities in district and state accountability systems has been targeted by policymakers as a critical element in the current push for educational reform (Geenen, Thurlow, & Ysseldyke, 1995; U.S. Department of Education, 1999; Ysseldyke, Thurlow, Algozzine, Shriner, & Gilman, 1993). This targeting is a result of documented corruption in accountability systems resulting from the over-exclusion of students, particularly students with disabilities (McDonnell, McLaughlin, & Morison, 1997; McGrew, Thurlow, Shriner, & Spiegel, 1992; Ysseldyke & Thurlow, 1994), as well as the associated increases in referrals to special education and rates of retention in grade (Allington & McGill-Franzen, 1992; Ysseldyke, Thurlow, McGrew, & Shriner, 1994; Zlatos, 1994).

The primary method for increasing the inclusion of students with disabilities in accountability systems is to increase their participation in regular district and state assessments. A number of strategies also have been suggested for increasing the participation of these students in large-scale assessment programs (see Elliott, Thurlow, & Ysseldyke, 1996; Thurlow, Elliott, & Ysseldyke, 1998; Thurlow, House, Boys, Scott, & Ysseldyke, 2000; Thurlow, Seyfarth, Scott, & Ysseldyke, 1997). The provision of assessment accommodations is a pivotal approach to addressing this issue (Elliott & Thurlow, 2000; Thurlow et al., 1997; Thurlow et al., 2000; Thurlow, Ysseldyke, & Silverstein, 1995; Ysseldyke, Thurlow, McGrew, & Shriner, 1994; Ysseldyke, Thurlow, McGrew, & Vanderwood, 1994). While the use of assessment accommodations is the most viable way to increase the participation of students with disabilities in accountability systems (Mazzeo, Carlson, Voekl, & Lutkus, 2000; Olsen & Goldstein, 1997), it is one of the more controversial aspects of current assessment discussions.

The controversy was heightened for a long time by the lack of a systematic and comprehensive research program to study the effects of accommodations on the psychometric characteristics of assessment results. This lack of research on accommodations has changed to some extent in recent years because of support from the U.S. Department of Education for research on a variety of issues related to the participation of students with disabilities in large-scale assessments. This federal support has come both from the Office of Special Education Programs (OSEP) and the Office of Educational Research and Improvement (OERI). In addition, some research has been conducted by the National Center for Education Statistics because of the relevance of this issue to the National Assessment of Educational Progress (NAEP). Additional research efforts have been supported more recently by some states and by some test publishers.

Prior to this new emphasis on accommodations research, there were two primary efforts to look at the effects of accommodations. In 1984, Laing and Farmer produced a report on issues pertaining to participation in the American College Testing Program (ACT) assessment by examinees with disabilities. The report summarized some information gathered from ACT records

from 1978-79 through 1982-83. Five groups of examinees were considered: students without disabilities and students with disabilities who took the exam in a standard administration, and students with visual impairments, hearing impairments, or motor disabilities (identified as including physical and learning disabilities) who took a nonstandard administration. Predictive validity was examined using first-year college grades as the criterion measure. It was reported that the prediction of first-year college GPA was about equally accurate for examinees without disabilities testing under standard conditions and examinees with visual disabilities under nonstandard testing conditions. Prediction was less for individuals with other disabilities (e.g., physical disabilities). Questions can be raised, however, about whether the college environment provided the necessary accommodations for their grades to align to their tests scores.

The Educational Testing Service (ETS), which was the other primary source of past research efforts on the effects of accommodations, examined more than just prediction of college scores. ETS conducted a series of studies on the comparability of standard and nonstandard versions of the Scholastic Aptitude Test (SAT) and the Graduate Record Examination (GRE) General Test. One such study examined item performance for students with disabilities who took the SAT with accommodations as compared to the group of students (assumed to be non-disabled) who took the SAT without any accommodations (Bennett, Rock, & Kaplan, 1987). When completing item analyses, the researchers condensed the items into clusters in order to reduce the statistical error. They found that two clusters on the mathematical scale were differentially difficult for visually impaired students taking the Braille version of the SAT, one cluster that included questions using graphics in a multiple choice format, and another that included miscellaneous multiple choice items. Additionally, it was found that the algebra comparison cluster was unexpectedly easy for students with learning disabilities taking the test via cassette, and for hearing-impaired students who took the regular exam with extended time. Finally, the researchers reported that when comparing items requiring differential amounts of reading, the hearing impaired students who took the regular exam with extended time found the non-reading cluster to be unexpectedly easy for them.

Some people might cite research by ETS and ACT as already providing answers that we can use; however, both ETS and ACT were looking at a limited number of accommodations, using non-representative samples of students (only those applying for entrance to postsecondary education institutions), and primarily focusing on predictive validity (Laing & Farmer, 1984; Willingham, Ragosta, Bennett, Braun, Rock, & Powers, 1988). Thus, these research efforts do not meet the current tremendous need for research on accommodations for state and district assessments.

There are many other isolated studies that are relevant to questions about the effects of accommodations. This far-reaching set of studies was summarized recently by Tindal and Fuchs (1999), who synthesized research on test changes. In addition to summarizing 115 studies, they

speak to the methodology of the studies. Noting that most of the research is "in a fragile position between program evaluation and quasi-experimental research" (p. 93), they identify several problems with much of the research conducted to date. They also strongly argue that in the end the research must "be validated with findings of an interaction between students with and without disabilities as they perform with and without the change" (p. 95). These types of findings require an experimental research design.

We need new research to answer questions about the validity of test results for students with a variety of disabilities, using a variety of accommodations, in district, state, and national assessments for which the purpose is to describe the status of student knowledge (Thurlow, Elliott, Ysseldyke, & Erickson, 1996). Investigating the effects of assessment accommodations on the accuracy and meaning of the resultant test scores is one of the most critical needs if the scores of students with disabilities are to be included in accountability systems. The research conducted thus far only begins to address this need and much of the research is inadequate.

In addition, the field also desperately needs research on the decision-making process relevant to accommodations. For some time now, it has been a strong suspicion that individuals making decisions about the specific accommodations a student needs have been doing so without an objective basis for the decisions (see Thurlow, Elliott, & Ysseldyke, 1998). These suspicions seem to be confirmed by data from those states that track the use of accommodations during state assessments (Thompson & Thurlow, 1999), in which the variation in numbers of students using accommodations is from 8% to 82%. Similarly, an empirical study now has demonstrated variance between teacher recommendations for accommodations versus those that actually boost students' scores on assessments (Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000). Research of this type has just started; much more quality research on the topic of accommodations decision-making is needed.

Definition of Terms

The term "accommodation," when used in relation to assessment, generally means some change that is made in the testing materials or administration of the test. There are numerous other terms that are used, among them "alteration," "modification," and "adaptation." The proliferation of different terms results in confusion; for this paper the term "accommodation" will be used. Accommodations can be divided into six categories (Thurlow, Elliott, & Ysseldyke, 1998), including accommodations that alter: (1) the setting in which the assessment is administered, (2) the timing of the assessment, (3) the scheduling of administration, (4) the presentation of the assessment, (5) the response that a student makes to the assessment, and (6) other kinds of changes. Table 1 presents examples of each of these types of accommodations. A common

Table 1. Examples of Six Types of Assessment Accommodations

Setting	**Presentation**
Study carrel	Repeat directions
Special lighting	Larger bubbles on multiple-choice questions
Separate room	Sign language presentation
Individualized or small group	Magnification device
Timing	**Response**
Extended time	Mark answers in test booklet
Frequent breaks	Use reference materials (e.g., dictionary)
Unlimited time	Word process writing sample
Scheduling	**Other**
Specific time of day	Special test preparation techniques
Subtests in different order	Out-of-level test

(Thurlow, Elliott, & Ysseldyke, 1998, p. 30)

"other" accommodation is the use of "out-of-level" testing. Some accommodations, such as out-of-level testing, are controversial while others have been accepted with little controversy (e.g., use of Braille versions).

Purpose of Assessment Accommodations

Assessment accommodations are intended to compensate for an individual's disability (i.e., level the playing field), not to give an individual with a disability an advantage over individuals without disabilities. Determining whether the playing field is equalized or biased is difficult. The observation that a student earns a higher score when using an accommodation than when not using an accommodation does not mean that the individual is receiving an unfair advantage over others. Thus, it is challenging to design research studies that adequately address questions about the technical fairness of accommodated assessments.

Some accommodations relate more closely to the construct being assessed than others, and when a student's disability requires the use of an accommodation that is closely related to the construct being assessed, the issues become very confusing and perhaps impossible to separate. For example, a student with a reading disability may require the assistance of a reader when participating in assessments. This seems logical when the focus of the assessment is mathematics or science, but becomes problematic when the focus of the assessment is reading. The assistance of a reader may be more acceptable when the focus of the assessment is reading comprehension rather than decoding skills.

As noted in the existing analyses of accommodations policies (Thurlow et al., 1997; Thurlow et al., 2000), there are several accommodations that are more controversial, including having an exam read to the student, extending test time, allowing the use of calculators, and using word processors with spelling and grammar checkers. These controversies become the core of legal actions when there are concerns about diplomas and whether they "have the same meaning for students who passed with and without accommodations" (Phillips & Millman, 1996, p. 1). Phillips and Millman also noted that there are similar concerns about the "fairness to low achieving students who may have been successful with an accommodation but do not qualify because they lack a diagnosed disability" (p. 2).

Little research on the effects of accommodations has been completed to date (see Fuchs et al., 2000; Thurlow, Hurley, Spicuzza, & El Sawaf, 1996; Thurlow et al. 1995; Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998), and until recently most of the research focused on college entrance exams and emphasized predictive validity (see Ragosta & Wendler, 1992; Willingham, Ragosta, Bennett, Braun, Rock, & Powers, 1988). Despite the renewed emphasis on conducting research on accommodations, many of the new research efforts still are struggling with how best to examine the effects of accommodations.

While it is possible to identify research designs that are appropriate for examining the effects of accommodations, practical constraints often reduce the feasibility of these research designs. There are constraints associated with district and state assessments, the settings that generally form the context within which applied research is conducted. Ethical considerations surrounding the withholding of accommodations and a variety of logistical constraints often make it impossible to apply the ideal research design.

Despite the difficulties associated with research on the effects of assessment accommodations, there is a critical need for good research to be conducted. Toward this end, it is useful to consider the issues that impinge on accommodations research and to identify the specific research questions that need to be answered.

Issues in Accommodations Research: Critical Research Questions

The psychometric/technical accommodation question that is most frequently asked is whether scores gathered under nonstandard conditions (i.e., with accommodations) can be combined with scores gathered under standard conditions. Combining both types of scores in aggregate reports assumes that test scores gathered under standard and nonstandard conditions are measuring the same abilities or constructs. Thus, a key issue in the use of assessment accommodations is validity or score comparability. Fundamental questions that must be addressed in accommodation research include:

- Are the scales underlying the individual items administered under standard and nonstandard conditions comparable (differential item functioning)? Or, can items administered under standard and nonstandard conditions be placed on the same measurement scale?

- Are the scores gathered under standard and nonstandard conditions measuring the same abilities or constructs (i.e., construct validity)?

- Do the scores gathered under nonstandard conditions correlate to the same degree with outcome criteria as do scores gathered under standard administration conditions (criterion-related validity)?

- Should a different cutoff score or standard be used for test scores gathered under nonstandard conditions when the scores are used to make important decisions about examinees?

To begin to address these kinds of issues, we need to think through good research designs. This process is not simple given the nature of disabilities and how these disabilities can interact with conducting research:

> A major limitation of collecting data on students with disabilities is the small samples typically available for specific combinations of disability and accommodation. Even when there are multiple students with the same disability, the degree of disability may vary markedly or there may be additional disabilities present that would limit valid generalizations.
>
> However, in a statewide program there may be a large enough population of students with learning disabilities to permit some useful data collection. For example, a subpopulation of students with specific learning disabilities in reading and no additional disabilities could be identified. To provide a consistent definition of reading disabilities, a specified range of standard score or regressed difference between ability and reading achievement could be used to define this group. (Phillips & Millman, 1996, p. 3)

To answer score comparability questions, appropriate analytic strategies and research designs must be applied to sufficiently large samples of students who are given the test under different administration conditions (i.e., with and without accommodations). A number of general data analytic strategies and research designs are needed. To answer questions about accommodation decision making, the research design issues are different. In this case, single subject research designs are important to consider.

General Analytic Strategies

Although there are many important issues and questions that need to be addressed in research on the effects of accommodations on test scores, some of the most central issues need to focus on the question: "Do accommodations change the nature of what is being measured by the test?" If the answer is "no," then the scores obtained under nonstandard conditions can be placed on the same measurement scale used for all students, and the scores can be aggregated and compared. To address these score comparability or validity questions, differential item functioning, factor analytic, and criterion-related data analytic strategies need to be used. A brief description of each of these strategies follows.

It is important to note, however, that this list of analytic strategies is illustrative—not exhaustive. For example, there are many other accommodation, research design, and sampling issues (e.g., accommodation or treatment integrity, randomly selected samples) that are not addressed in this conceptual paper. Furthermore, differences between and within district or state testing programs most likely will require unique variations and modifications of these strategies.

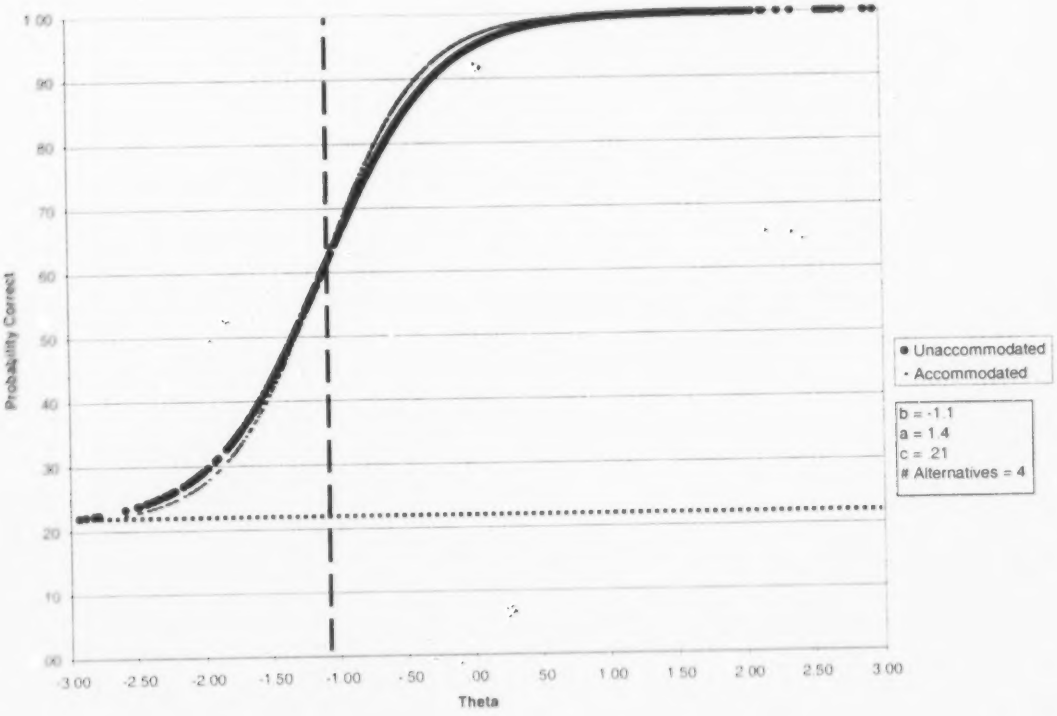
Item Response Theory (IRT)

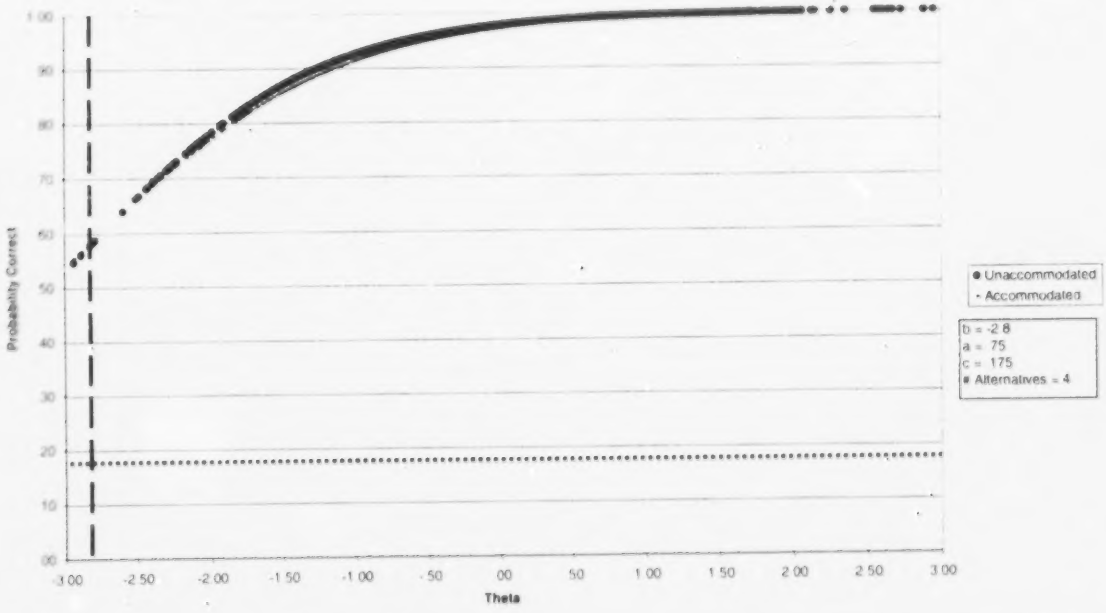
Item Response Theory (IRT) strategies are recommended for evaluating the extent to which the abilities measured by the individual items of a test are changed substantially as a result of an accommodation (i.e., investigating differential item functioning). These strategies are explained using Figures 1 and 2 as examples that illustrate the application of IRT methods to four items from the 1995 NAEP Field Test.

In each figure, the x-axis represents the ability or trait being measured by the items in the test (in this case, mathematical ability). The scale is a continuum from less (left end) to more (right end) mathematical ability. The y-axis represents the probability of success on the item in question. By applying IRT procedures, an item characteristic curve (ICC) is developed for each item that visually represents the probability of success on that item (y-axis) as a function of ability (x-axis).

The two graphs in Figure 1 portray the ICCs for two mathematics items. The solid ICC lines represent the ICC obtained for the two items in the general population (i.e., how the items "behave" for most subjects who responded to the items under standard test conditions). Also plotted on the two graphs are the ICCs for the same items when given to students with disabilities under accommodated conditions. These curves are represented by the small dots. What one hopes to find, and what is represented in the two graphs in Figure 1, is a situation where the ICCs for standard and accommodated administrations are almost identical. A visual review of

Figure 1. Graphs Showing Similar Item Characteristic Curves for Standard and Accommodated Administrations





BEST COPY AVAILABLE

the two sets of ICCs for the two graphs shows that the ICCs are indeed similar. This means that the items appear to be "behaving" similarly regardless of whether they were administered with or without accommodations. Thus, they appear to be measuring the same trait or ability, and therefore can be placed on the same measurement scale.

Figure 2 presents a contrasting finding. In both graphs, the ICCs for the standard and accommodated administrations are dramatically different. ICC plots such as these, and their associated empirical fit indicators, suggest that when the items are administered under nonstandard conditions, they "behave differently." That is, the same items are not measuring the same trait or ability when an accommodation is introduced. The empirical relationship between the probability of success on these items and the traits or abilities being measured has been altered by the use of an accommodation. As a result, test scores generated from the combination of a large number of these "misbehaving" items cannot be placed on the same measurement scale as scores based on a combination of items administered under standard conditions.

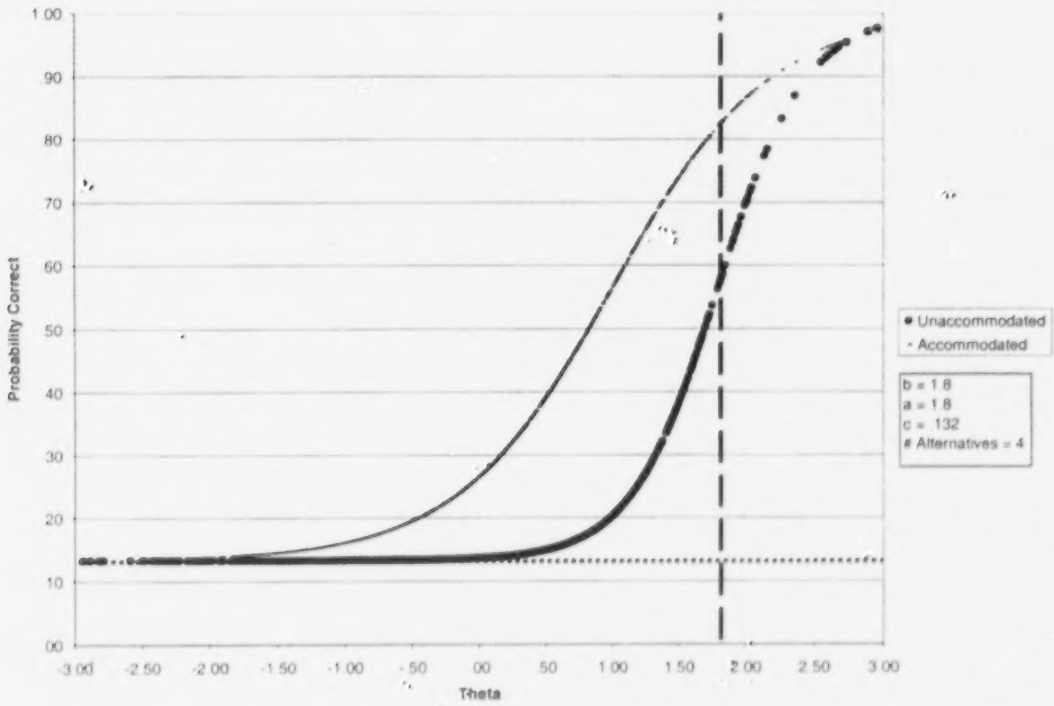
Factor Analysis

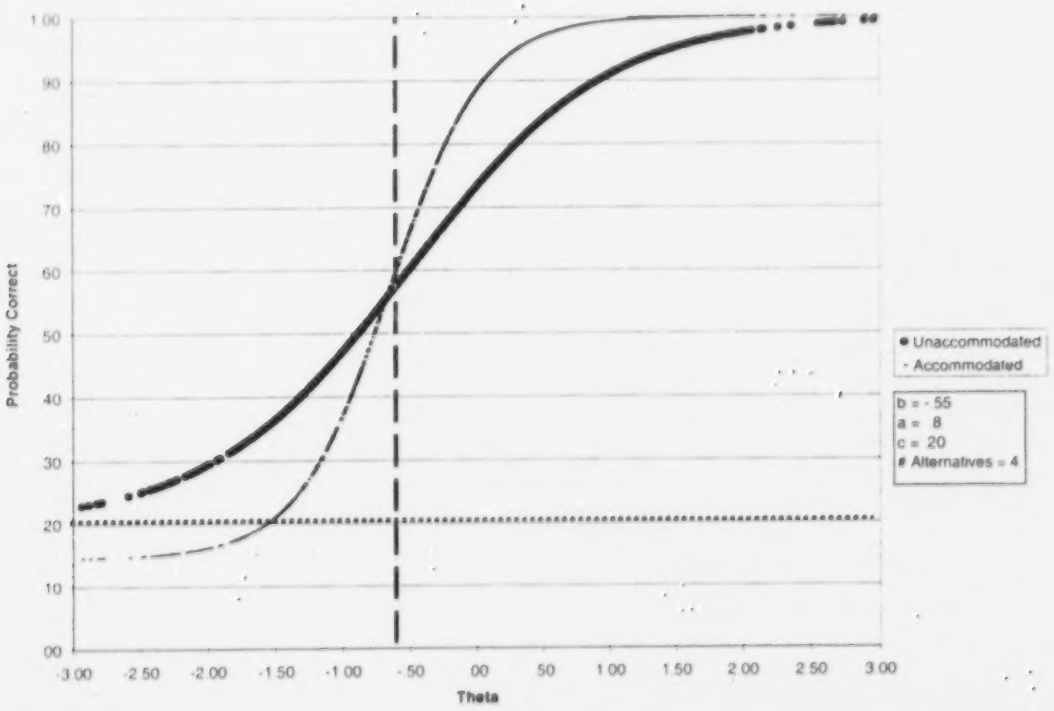
Factor analytic strategies are important for evaluating the construct validity of tests. These procedures help determine whether the underlying dimensions or constructs measured by a test are the same when administered under standard and nonstandard conditions. The two diagrams shown in Figure 3 illustrate the essence of these types of analyses.

The rectangles in the figure represent sub-scales A-F from a math test. Each subscale is constructed from a combination of math items that together measure a mathematics subskill. When given a set of variables or subscales (in this example, six math subscales), factor analytic procedures help determine the number of broader dimensions, factors, or constructs that account for the shared abilities of the subscales. In the first factor model, the six subscales (A-F) were found to be indicators of one general construct of math (viz., General Math). The circle represents this factor or construct. Assume that this single or general factor model was found when the math tests were administered to the general population (without accommodations) and the data were factor analyzed.

Next, assume that the same six subscales were administered under accommodated conditions. If the accommodations do not change the nature of the construct being measured, then the application of factor analytic procedures to the data from this sample should result in generally the same factor structure. That is, the construct being measured by the test under nonstandard conditions would be the same if a similar single or General Math factor was found to best represent the relationship between subscales A-F. Alternatively, if the accommodations changed the nature of the construct being measured, a different factor structure might emerge.

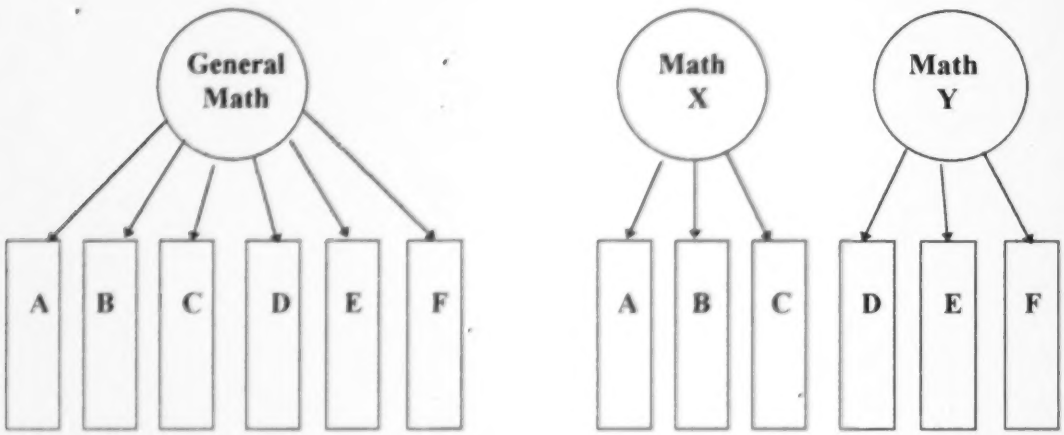
Figure 2. Graphs Showing Different Item Characteristic Curves for Standard and Accommodated Administrations





BEST COPY AVAILABLE 18

Figure 3. Factor Analysis Models Indicating Different Factor Structures for Standard and Accommodated Administrations



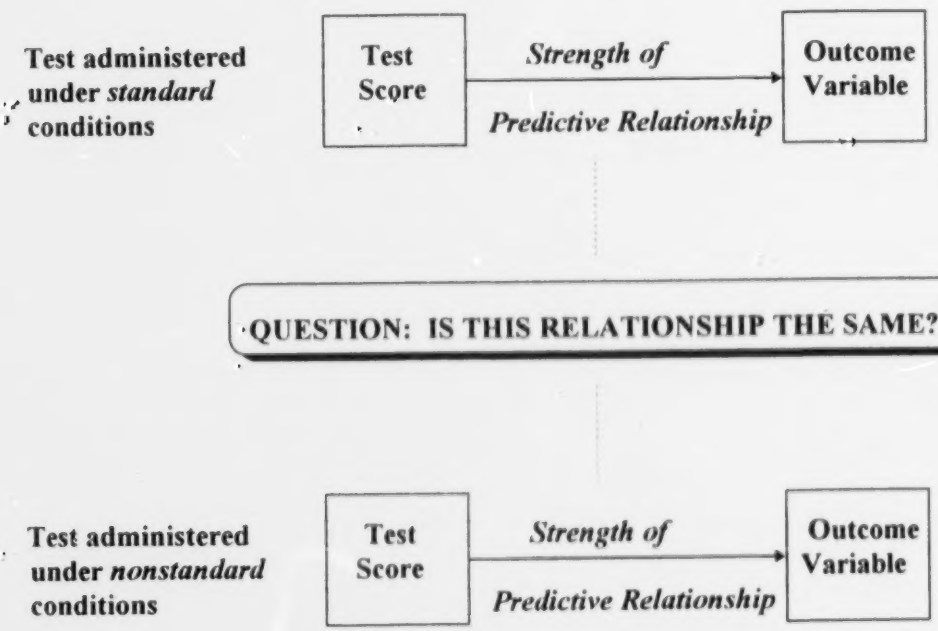
The second model displays factor analytic results that suggest that the structure or dimensions being measured by subscales A-F under accommodated conditions are best explained by two different broad math factors (Math X and Math Y). This finding, together with the finding of the General Math factor model in the general population, would indicate that this specific collection of math subscales is not measuring the same constructs under standard and accommodated conditions. Besides exploratory factor analysis, confirmatory factor analysis (LISREL) procedures can be used in these types of analyses. Confirmatory procedures are particularly well suited to evaluating the extent to which the constructs measured by a collection of variables or tests are similar (i.e., invariant) across different samples and conditions.

Criterion-Related Analyses

Criterion-related analytic strategies are needed to investigate the extent to which accommodated test administrations change the relationship between test scores and other criteria. These procedures can help evaluate whether the criterion-related validity (often referred to as predictive validity) of a test is similar for different samples or for different versions of the same test (i.e., standard and accommodated test administrations). If a test is used to make predictions about a person's performance on an important outcome criterion (e.g., potential success in college, mastery of a domain of skills), it is important to know whether the relationship that exists between the test score(s) (i.e., the predictor) and the important outcome criteria changes when the test is administered under accommodated conditions. That is, can prediction and classification decisions about a person be made with a similar degree of confidence for test scores administered under standard and accommodated conditions?

Although the specific data analytic method may vary depending on the nature of the predictor and outcome variables (e.g., correlation, multiple regression, classification agreement), most criterion-related analytic strategies are concerned with addressing the question represented in Figure 4, "Is this relationship the same?"

Figure 4. Representation of Criterion-Related Analytic Strategies



Group Research Design Considerations

To use the general analytic strategies described above, research designs must be employed that meet certain characteristics. This section presents general design considerations for sampling methods and sampling size in group-based accommodations research. These are presented to provide an idea of design considerations that may be required to conduct research on the effects of accommodations on test scores.

Sampling

Sampling issues are very complex and cannot be treated in detail in this paper. Ideally, the samples in each design matrix cell would be randomly selected from the appropriate population

(e.g., Group 1 and Group 2, both randomly selected from all students with the particular characteristic or accommodation need being targeted in the larger population of interest).

Size

Another important consideration is the size of the sample in each design cell. The general analytic strategies described above (factor analysis and IRT, in particular) require relatively large samples to obtain stable statistical estimates. Many measurement specialists would recommend sample sizes as large as 500 for each cell in each design matrix for IRT analyses. However, given the practical constraints of applied research, and the small number of students with disabilities who take tests with accommodations, smaller sample sizes are more realistic. We suggest that, at a minimum, 200 subjects per subsample (i.e., each cell in each design matrix) should be used for applied research employing the general data analytic strategies outlined in this paper.

Group Research Designs

The four general group research designs presented in this section are ordered from the most optimal (Design # 1) to the least optimal (Design # 4). For illustrative purposes, only one type of accommodation group (e.g., students with reading difficulties or who need a specific type of accommodation) is presented in each design. Additional groups, or students with other characteristics (e.g., limited English proficiency), with parallel information in each cell, could be added to the design matrices. In addition, we have presented the simple version of each design. Any of the designs could be made more sophisticated by counterbalancing not only form of the test, but also order in which forms are presented, and so on. The designs that we present can be modified in many ways. It is also important to note that we do not define accommodation groups by disability category since category of disability does *not* define the need for accommodations. Nevertheless, it generally is helpful to select subjects meeting a specific criterion (e.g., reading problem identified by test score) from within a single disability category (e.g., learning disability) so that other complicating characteristics (e.g., visual disability) are less likely to complicate findings.

Design 1

This design allows for the examination of the comparability of scores as a function of the presence/absence of a characteristic, the use of an accommodation, and the interaction of these two factors. The design requires equivalent forms (A & B) of the test. The effect of test order is

controlled by counterbalancing the administration of forms A and B. Thus, Design 1 requires subjects who are willing to take two versions (with and without accommodations) of the same test. Subjects with and without disabilities who take the test without the accommodations could be drawn from the general testing population. Their scores could be randomly selected from the total test sample of all students who regularly take the version of Forms A and B. This design does not require that the samples from the two respective groups (Disability groups 1 and 2 and Non-disability groups 1 and 2) be exactly similar (i.e., matched) in important characteristics. Design 1 is illustrated in Table 2.

Table 2. Design 1: Comparability of Scores as a Function of the Presence/Absence of a Disability

	Disability Group 1*	Disability Group 2*	Non-Disability Group 1	Non-Disability Group 2
With Accommodation	Test Form A	Test Form B	Test Form A	Test Form B
Without Accommodation	Test Form B	Test Form A	Test Form B	Test Form A

* Disability Groups 1 and 2 are students with a common characteristic (e.g., students with reading problems) or who have the same accommodation need (e.g., Braille edition).

An example of a study that used Design 1 is a recent multi-state study supported by the Technical Guidelines for Performance Assessment project, which received funding from the U.S. Department of Education, Office of Educational Research and Improvement (OERI). In this study, groups of students with reading disabilities and students without any special education designation were administered the equivalent of a state test using a videotape presentation of a math test and the test administered under typical conditions. Two forms of the test were given to both groups, in counterbalanced order, to begin to sort out the effects of the change in test administration procedures.

Design 2

Design 2 also allows for the examination of the comparability of scores as a function of the presence/absence of a disability-related need, the use of an accommodation, and the interaction of these two factors. It is different from Design 1 in that this design requires that the respective samples of the students with disabilities groups (Groups 1 and 2) and students without disabilities (Groups 1 and 2) be equivalent in important characteristics (e.g., matched samples). If they are not, it is impossible to determine whether any differences between the score characteristics of

the respective groups (Disability Group 1 vs. Disability Group 2; Non-Disability Group 1 vs. Non-Disability Group 2) are due to the effects of the accommodations, or are attributable to differences in sample characteristics. Design 2 does not require equivalent forms (A & B) of the test. Subjects with and without disabilities who take the test without accommodations can be drawn from the general testing population. Their scores can be randomly selected from the total test sample of all students who regularly take versions of Form A (see Table 3).

Table 3. Design 2: Comparability of Scores as a Function of the Presence/Absence of a Disability

	Disability Group 1*	**Disability Group 2***	**Non-Disability Group 1**	**Non-Disability Group 2**
With Accommodation	Test Form A		Test Form A	
Without Accommodation		Test Form A		Test Form A

* Disability Groups 1 and 2 are students with a common characteristic (e.g., students with reading problems) or who have the same accommodation need (e.g., Braille edition).

A version of this design was used by Tindal, Hollenbeck, Heath, & Almond (1998), who had students take a statewide writing test that required them to write a composition. The students were allowed use of either paper and pencil or a computer over the three days devoted to writing the composition. In this study there was the additional condition that students could use the computer to (1) compose on all three days, (2) compose on the computer only the last day, or (3) compose with a spell-checker available. The compositions were compared on six traits (ideas-content, organization, voice, word choice, sentence fluency, conventions).

Phillips and Millman (1996) noted that beyond the selection of comparable students, there are additional concerns, such as the standardization of equipment and ensuring that students had adequate training in word processing:

> Standardization of equipment is an issue because the study would probably rely on the use of computer equipment already present in the schools. Because different software programs offer a variety of options, it would be necessary to develop a list of permissible equipment and software, which is judged to provide the same basic features and ease of use. Spell-check, thesaurus and editing functions should be comparable. Finally, each student should be thoroughly familiar with the hardware and software to be used during testing and should have had sufficient practice time to develop facility with the software (p. 4).

Design 3

Design 3 allows for the examination of score comparability as a function of accommodation use for only one disability group. This design requires the assumption (based on prior research) that the scores of subjects with disabilities who take the test without the accommodation are comparable to the scores of subjects without disabilities who take the test without the accommodation. It requires equivalent forms (A & B) of the test, and controls for the effect of test order by counterbalancing the administration of forms A and B. Design 3 also requires subjects who are willing to take two versions (with and without accommodations) of the same test. Subjects with disabilities who take the test without the accommodation can be drawn from the general testing population. Their scores can be randomly selected from the total test sample of all students with disabilities who regularly take the versions of Form A and B. Finally, Design 3 does not require that the two respective samples (Groups 1 and 2) be exactly similar (i.e., matched) in important characteristics. This design is illustrated in Table 4.

Table 4. Design 3: Examination of the Comparability of Scores as a Function of the Use of an Accommodation for a Single Disability

	Disability Group 1*	Disability Group 2*
With Accommodations	Test Form A	Test Form B
Without Accommodations	Test Form B	Test Form A

* Disability Groups 1 and 2 are students with a common characteristic (e.g., students with reading problems) or who have the same accommodation need (e.g., Braille edition).

An example of a study that used something like Design 3 is one conducted by Tindal, Heath, Hollenbeck, Almond, and Harniss (1998). They had students complete reading and math multiple choice tests by either filling in the standard bubble sheets or by marking on the test booklet.

Design 4

Design 4 allows for the examination of the comparability of scores as a function of the use of an accommodation for subjects with disabilities only. This design requires the assumption (based on prior research) that the scores for students with disabilities who take the test without the accommodation are comparable to those for regular education students who take the test without the accommodation. It also requires that the respective subjects with disabilities be equivalent in important characteristics (e.g., matched samples). If not, it is impossible to determine whether

any differences in score characteristics between the respective groups are due to the effect of the accommodation or are attributable to differences in sample characteristics. Design 4 does not require equivalent forms (A & B) of the test (see Table 5). Subjects who take the test without the accommodation could be drawn from the general testing population. Their scores can be randomly selected from the total sample of all students with disabilities who regularly take versions of Form A.

A study using this design could take place during an actual large-scale testing session. As an example, Design 4 could be used in pre-selecting students with disabilities who would be participating in a large-scale assessment. Students could be matched by nature of disability (e.g., reading disability) and other important factors. Students in Group 1 would have the test read aloud while students in Group 2 would read the test to themselves. Scores could be compared for evidence of differences between groups. Of course, when using this design, researchers must ensure that students would not be denied accommodations they need, especially if the results would be used for high-stakes decision making.

Table 5. Design 4: Examination of the Comparability of Scores as a Function of the Use of an Accommodation for Subjects with Disabilities

	Disability Group 1*	Disability Group 2*
With Accommodation	Test Form A	
Without Accommodation		Test Form A

* Disability Groups 1 and 2 are students with a common characteristic (e.g., students with reading problems) or who have the same accommodation need (e.g., Braille edition).

Single Subject Research Designs

Research on accommodations need not be limited to group comparison designs. There are other designs that provide important information that will inform both practice and future research. Single-subject research designs constitute a viable and emerging approach to accommodations research. The single subject design fits nicely within the realm of research on the participation of students with disabilities in assessments. The purpose of using a single subject research design is to first determine whether an accommodation is effective for an individual student, and also to search out the reason for the effects.

Single subject research designs provide a more specific way of understanding functional relationships between environmental events and behavior. In this definition, emphasis is placed

on the function of behavior: How does behavior change contingent on specific environmental events? This question usually is answered by clearly describing behaviors; identifying events, times, or situations that predict the occurrence of behaviors; identifying consequences of the behaviors; developing summary hypotheses of possible explanations for the occurrence of the behavior; and using various observational systems to determine the co-occurrence of behaviors either in specific settings or in the presence of specific discriminative stimuli; or to follow changes in behavior as a function of specific consequences.

For example, imagine a student with a learning disability who exhibits several behavioral problems (low attention span, quick and impulsive responding, high rates of "fidgeting," and verbal outbursts) in addition to severe reading skill deficits. Not only would this student have a difficult time taking a large-scale test in a group setting in one session, but others taking the test may be hindered from performing optimally. Let us assume that an adequate assessment had been done initially and that the student is receiving services in reading, with some form of curriculum-based measurement for monitoring progress; that is, at least twice each week the student completes a one-minute oral reading fluency probe in which a passage of text is read aloud and the number of words read correctly is counted and graphed. A review of the IEP also indicates that in the general education class, a student is allowed to take tests individually and over multiple sessions with frequent breaks. Furthermore, the directions are highlighted to emphasize how a student is to respond. In this situation, the kind of testing done in either special or general education is quite different from most statewide multiple-choice reading tests. When the statewide test is administered it would seem sensible for most IEP teams to think that some of these accommodations (individual and brief sessions with highlighted directions) should be implemented as a matter of course without conducting any research.

The phrase "single subject design" provides a broad description for a host of designs that differ considerably from each other but all have one element in common: individual behavior is monitored over time (using repeated measurement) along with the systematic introduction of various "treatments," the goal of which is to determine the "cause" of a specific behavior that has been operationally defined. The major reason for selecting a single subject research design is the need to understand a specific treatment effect on an individual's behavior. Although group designs can reflect the effects of treatments on groups of students, they represent generalizations that must be further corroborated with specific individuals. The reason for this qualification is that rarely are treatments equally effective for all individuals and when an effect is found, it simply means that the treatment was more effective for most individuals. It does not mean that it was effective for all individuals or that, even with those who were positively influenced by the treatment, the levels of effect were equal. Another way to describe this outcome when analyzing treatment effects with traditional statistics is that the treatment showed more variation *between* its presence and absence than existed *within* either of these conditions.

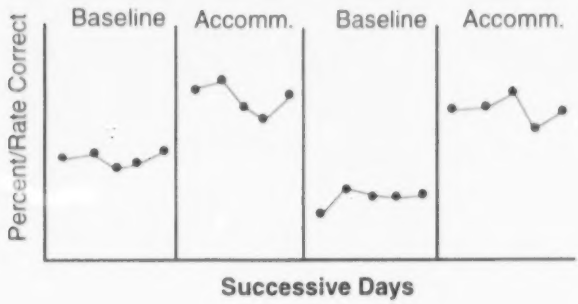
Another reason for using a single subject design is its practicality when conducting exploratory research. Rather than developing interventions that have to be implemented with a large number of individuals, it is possible to get a general indication of the potential for a specific treatment by systematically studying it with several individuals.

Finally, single subject designs provide a strategy for evaluating complex cause and effect relationships under changing conditions and provide for more precise isolation of essential treatment elements. Five types of designs are considered in this section: the basic withdrawal-reversal design, multiple baseline designs, multiple probe designs, changing criterion designs, and comparative designs.

Withdrawal-Reversal Design

Withdrawal-reversal design presents students with a series of alternating conditions in which the first baseline is taken, then treatments are implemented, followed by a return to baseline conditions, and a second return to a treatment condition (see Figure 5). At times the return to treatments is further extended with variations on the initial treatment. This design presents a great advantage over either a baseline only design (A), a treatment only design (B), or the simple combination of a baseline-treatment design (AB), and reflects the minimum design for quasi-experimental analysis of cause-effect. Neither of these more elemental designs allow any understanding of potential causes and both have many threats to internal validity.

Figure 5. Graphic Depiction of a Withdrawal/Reversal Design



Because of the unique features of academic skills in which learning occurs and with the unique characteristic of reversing or withdrawing the treatment, this design is particularly appropriate for studying test accommodations that really reflect conditions of measurement separate from the construct being measured. If a skill level can be established at a certain rate or level of accuracy or production with no accommodations, and then when specific changes are made in the way the test is administered or taken and performance is improved, the accommodation is

partially vindicated. This outcome presumes that the levels (rates) of access skills in the presence of certain test conditions during baseline are sufficiently low to be certain that the only variable limiting performance is the testing conditions or related access skills. Two examples of Withdrawal-Reversal Designs follow.

In measuring math performance, it may be important to eliminate reading as an access skill; the accommodation research then would focus on the effect of reading the math test to students. It would be necessary, however, to measure reading skill (i.e., oral reading fluency) before or during the measurement of math. If reading performance levels are low and math performance is improved when the test is read to the student, the empirical basis for the accommodation is strengthened. If reading proficiency is high, it is unlikely to change math performance.

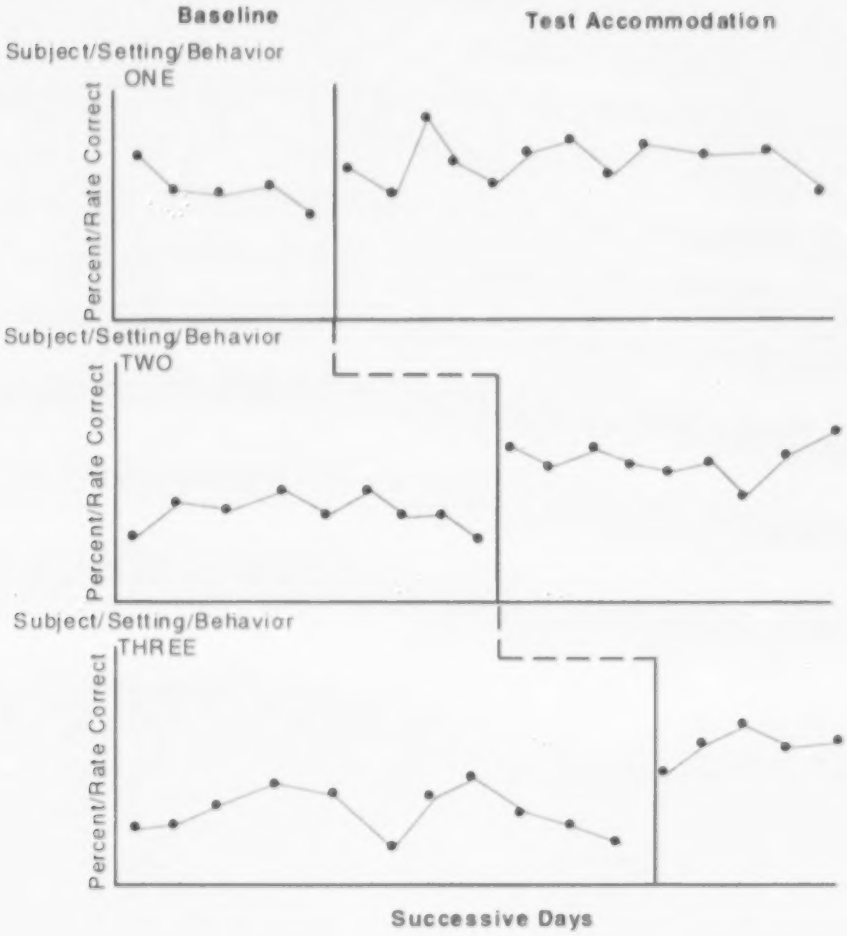
Accommodations in scheduling or setting need to be used when performance is low (not just on the skill being measured but on the access skills being used as part of the testing conditions). For example, if during the standard administration of a test, a student is observed working for only a few minutes (low percentage of the testing time) and then, under the accommodated test administration condition (use of frequent breaks), a student is observed working a high percentage of the time; the testing conditions appear relevant in influencing performance. If a student also exhibited higher performance on the test when the multiple breaks were used than when the test was administered in larger blocks of time, the accommodation is justified and an empirical basis for its use is present.

Multiple Baseline Designs

Multiple baseline designs actually refer to the lagged introduction of the treatment condition across subjects, conditions (settings), or behaviors. The reason for the name is that each of these levels of manipulation has a baseline with typically at least three different replications across subjects, conditions, or behaviors (see Figure 6). This design is particularly useful when it is not possible to remove a treatment (either because it resulted in a skill, which could not return to baseline conditions to show the relative effects of the treatment, or because of ethical reasons). In lagging the treatment and determining whether the levels (or rates) do not change until the treatment is introduced, this design provides partial validation of the treatment as the cause of any behavior changes.

In addition to the requirement of both comparability and independence across the subjects, conditions (settings), or behaviors, this design assumes that delayed access to treatments in later baselines is not a problem. For some subjects or with some conditions and behaviors, an extended baseline needs to be conducted in such a manner that other factors do not enter into the outcomes (like frustration, fatigue, or vigilance on the part of the subject). In addition, an extended baseline also implies comparability in the conditions in which no changes occur in the

Figure 6. Graphic Depiction of Multiple Baseline Design



presence of the treatment or the collection of the outcome data (instrumentation and testing) during the entire phase; otherwise, any of the previously noted threats to validity may be present.

A multiple baseline design across settings may be functional in sorting out various scheduling accommodations in different subject areas having a common response demand. For example, in many discipline-specific tests (math, science, and social sciences), students may be given a problem to solve in which they write their answers. For students with individual needs in writing who cannot write as proficiently or fluently as other students, it may be important to break the test sessions up into smaller time periods (e.g., three 15-minute periods instead of one 45-minute period). To test this accommodation, a student would take the test in two areas (e.g., geography and economics) initially using the standard time twice each week for two weeks, then take the test daily in 10-minute periods in geography for two weeks while the economics problem-solving task is completed using the standard time. After two weeks, this test in economics

is then administered in five 10-minute periods for two weeks. As can be seen from this example, the problem with this design is the extended baseline (in economics) for four weeks. Another problem is the sheer length of time overall in this example, a total of six weeks.

Multiple Probe Designs

In multiple probe designs, the baseline condition is prolonged and only sample probes are taken to ascertain the levels (rates) of behavior (see Figure 7). The major reason to use this design is that an extended baseline may be completely unnecessary once it is documented that the performance levels are low. It is important to both establish this low performance level initially and then again just before the intervention is implemented. If multiple probes are taken just before the intervention only (when both baseline and treatment conditions are being implemented across the subjects, conditions, or behaviors), the data display is less convincing in documenting that the changes are concurrent with the introduction of the treatment. Levels of performance are never available for comparing subjects, conditions, or behaviors under a common baseline condition.

Changing Criterion Designs

The critical element of changing criterion designs is the systematic introduction of a criterion level of performance over successive phases so that the behavior is essentially shaped into a final level, with each change in behavior occurring concurrent with the change in criterion (see Figure 8). Experimental control is established by the simultaneous co-occurrence of both. In this design, successive levels of the criterion are changed only upon attainment of previous levels.

The following example illustrates the use of a changing criteria design. For a student with an attention deficit, poor performance may be a function of not attending to the problems and working only in brief periods. If the student is trained to remain attentive to a read aloud condition using a specific reinforcement schedule, then a test may be more appropriately used to assess academic skill (possibly math or other content areas in which reading should not limit performance). In this example, note that the accommodation includes both a behavioral skill (attending) as well an access skill (e.g., reading). With a changing criterion design, this accommodation may be investigated by successively increasing the length of time in which eye contact is made with the person reading the test. To exhibit experimental control, the researcher would systematically manipulate the reinforcement after different amounts of time. Concurrently, performance would be tracked on an outcome measure (e.g., math test) to determine if attending has an influence.

Figure 7. Graphic Depiction of a Multiple Probe Design

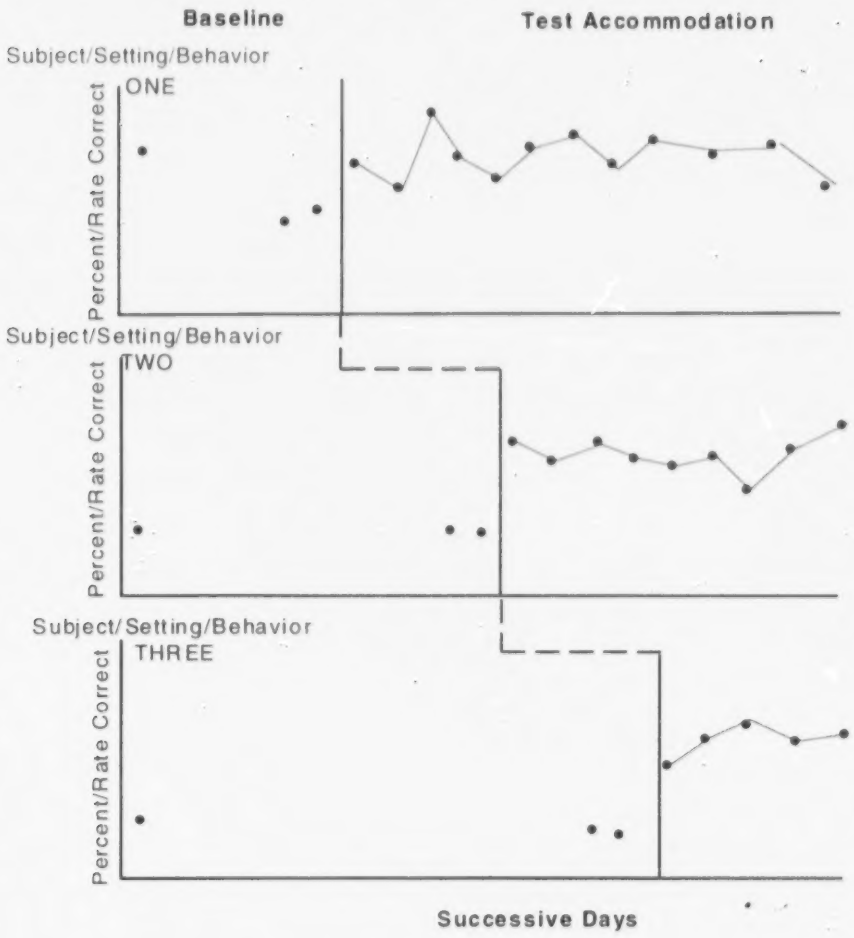
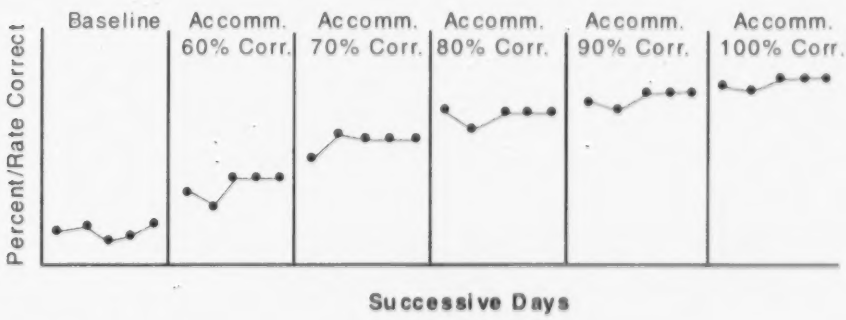


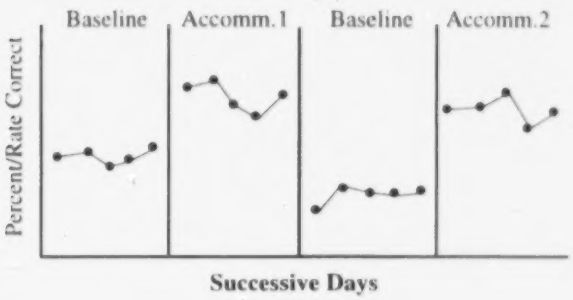
Figure 8. Graphic Depiction of a Changing Criterion Design: Increase Only



Comparative Designs

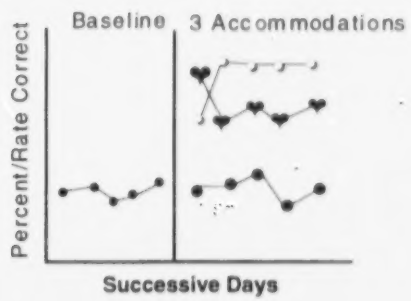
In comparative designs, different treatments are considered using any one of several strategies. For example in a multi-treatment design, successive phases highlight the distinctions between a variety of treatments and baseline (e.g., A-B-A-C-A-B-C or A-B-A-C). When making comparisons across successive phases, it is important to consider the sequence so that each phase is preceded and followed by every other phase in a balanced manner. In the example noted parenthetically, the alternate phases of B, C, and BC all follow a baseline, however, C and BC are confounded by being sequenced in that particular order. In some instances, this building of treatments in a sequence is unavoidable. In Figure 9, the sequence is depicted with two different treatments sequenced after each baseline in a multi-treatment design.

Figure 9. Graphic Depiction of a Multi-treatment Design



In contrast, in the alternating treatment or multi-element design, the treatment and control conditions are presented in a random order (counterbalanced within a session) so that successive days (or sessions within days) contain an unordered sequence (see Figure 10). This design relies on stimulus discrimination, allowing the subject to identify the conditions, and depends on a treatment being readily implemented and removed in a more quick fashion than a reversal-withdrawal design.

Figure 10. Graphic Depiction of an Alternating Treatment Design



An accommodation study may involve the use of an assistive device (e.g., a calculator or Franklin Speller) during the first accommodation and then a prompt to use the device in the second accommodation intervention; or in many instances, both a social skill and an administration assist may be needed and in that order. For example, a student may need to first be taught to work independently for 30 minutes using a reinforcement system and then a response accommodation might be implemented with various kinds of assistive devices (e.g., special computer keyboard).

Using another example, a student may be assisted in a reading test by having two forms of the test administered: (a) one form includes the passage written with one sentence per line, multiple choice items listed below, and a bubble sheet; and (b) one form has the passage presented in a standard form but allows a student to mark the multiple choice items in the booklet instead of on the bubble sheet. To use an alternating treatment design, a student would receive passages randomly with either of these two accommodations and directed to respond accordingly. The advantages of this design include the efficiency in comparing several treatments concurrently, the capacity to dismantle essential components of a treatment quickly, as well as the lack of reversal needed to understand the relative effects of a treatment.

Conclusions and Recommendations

A number of group and single subject research designs are available for understanding not only the effects of an intervention, but the degree to which critical variables help explain why the effect occurred. It is this latter issue, the need for an explanation or cause of an effect that makes the design a critical part of the research effort. To the degree that the experimental situation controls all possible threats to internal and external validity and the design is appropriately implemented, inferences can be made about both the effect and the reason for it. Of course, with a careful description of the students participating in the study and replications using different designs, it also is possible to begin establishing the ecological or external validity of the findings.

The difficulties that arise when conducting research on accommodations are not insurmountable, particularly when the research is to be done within the framework of actual assessments. They do, however, involve (1) stepping back to examine the major issues to be addressed, and (2) accepting some designs that may not give pure group comparison data but will give a good sense of what research questions are important to ask next.

In discussions at the National Center on Educational Outcomes, we have returned repeatedly to certain recommendations that we believe should guide thinking about research designs. These recommendations are listed below.

1. Focus on the accommodation(s) of most interest, either because they are the most controversial or because the largest number of students use them.

2. Focus on those students who comprise the largest part of the population with disabilities, either in a categorical sense (i.e., learning disabilities, speech and language disabilities, emotional disabilities, etc.) or in a severity-by-typological sense (e.g., students with average intelligence but low academic achievement, students with mild disabilities of any kind, etc.). Most often the target group will be students with learning disabilities or students with limited English proficiency.

3. At least one comparison group must have no evidence of disabilities, including poor academic achievement. Since the accuracy of defining which students are "eligible" for special education is unclear, and since there is some evidence of considerable overlap in students with learning disabilities and those who are low-achieving, for example, a comparison group of low-achieving students will produce unclear results. This is not to say that such a group would not be interesting to add to a design that included students with disabilities and students clearly without disabilities.

4. There must be a plan for the collection of other measures, ones that will help clarify findings. For example, additional measures might be used to assess students' skills that are related to the use of accommodations. For example, a study of the use of dictionaries would benefit from a brief assessment of the students' skills in using dictionaries, even if the students were trained to use the dictionary before the study began.

By following these general recommendations, the resulting research designs will have the greatest impact possible. Generally, research focused on students with low incidence disabilities or on accommodations that are used infrequently (or that are not available) will need to be conducted under more laboratory-type conditions, or through the aggregation of subsample data across several states.

In this paper, we addressed a number of group and single subject research designs in which test accommodations can be better understood. Although the paper was aimed at the researchers, test developers, and others who are using the findings of research, at the very least, teachers and others involved in serving students with disabilities should learn how accommodation decisions can be made from an empirical basis. Therefore, whether actually conducting research on accommodations or using the logic to make decisions, many educators should be able to use this information.

Still, there is an additional recommendation that requires consideration—the need for programmatic research on accommodations. Isolated studies probably will never completely answer the questions about accommodations that are plaguing the field. There is a dramatic

need for a program of research, with one question followed by additional clarifying questions, perhaps with different procedures used to ask slightly different questions each time. Even with the best research designs, we probably will not get nice answers unless we have a program of research to follow up with additional questions.

With the best research designs and good programmatic research, it is likely that we will, for some accommodations, be pushed to ask questions beyond the scope of accommodations research. At some point, we will have to answer the difficult questions about what we are really testing, and whether what we are testing is what we really should be testing.

References

Allington, R. L., & McGill-Franzen, A. (1992). Unintended effects of reform in New York. *Educational Policy, 4,* 397-414.

Bennett, R. E., Rock, D. A., & Kaplan, B. A. (1987). SAT differential item performance for nine handicapped groups. *Journal of Educational Measurement, 24* (1), 44-55.

Elliott, J. L., & Thurlow, M. L. (2000). *Improving test performance of students with disabilities in district and state assessments.* Thousand Oaks, CA: Corwin Press.

Elliott, J., Thurlow, M., & Ysseldyke, J. (1996). *Assessment guidelines that maximize the participation of students with disabilities in large-scale assessments: Characteristics and considerations* (Synthesis Report 25). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C., & Karns, K. (2000). Supplementing teacher judgments about test accommodations with objective data sources. *School Psychology Review, 29* (1), 65-85.

Geenen, K., Thurlow, M., & Ysseldyke, J. (1995). *A disability perspective on five years of education reform* (Synthesis Report 22). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Laing, J. & Farmer, M. (1984). *Use of the ACT assessment by examinees with disabilities* (Research Report 84). Iowa City, IA: American College Testing Program.

Mazzeo, J., Carlson, J. E., Voekl, K. E., & Lutkus, A. D. (2000). *Increasing the participation of special needs students in NAEP: A report on 1996 NAEP research activities* (NCES Statistical Analysis Report 2000-473). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

McDonnell, L. M., McLaughlin, M. J., & Morison, P. (Eds.) (1997). *Educating one and all: Students with disabilities and standards-based reform.* Washington DC: National Academy Press.

McGrew, K. S., Thurlow, M. L., Shriner, J. G., & Spiegel, A. N. (1992). *Inclusion of students with disabilities in national and state data collection programs* (Technical Report 2). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Messick, S. (1989b). Validity, In R. L. Linn (Ed.) *Educational Measurement - Third Edition* (pp. 13-104). New York: Macmillan.

National Research Council. (1999). *Testing, teaching, and learning* (R. F. Elmore & R. Rothman, editors). Washington, DC: National Academy Press.

Phillips, S. E., & Millman, J. (1996). *A design for assessing a reading aloud accommodation.* Unpublished manuscript.

Ragosta, M., & Wendler, C. (1992). *Eligibility issues and comparable time limits for disabled and nondisabled SAT examinees* (Report No. 92-95). New York, N.Y.: College Entrance Examination Board. (ERIC Doc. Rep. Service No. ED 349 337).

Thurlow, M. L., Elliott, J. L. Erickson, R., & Ysseldyke, J. E. (1996). *Tough questions about accountability systems and students with disabilities* (Synthesis Report 24). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L., Elliott, J. L., & Ysseldyke, J. E. (1998). *Testing students with disabilities: Practical strategies for complying with district and state requirements.* Thousand Oaks, CA: Corwin Press.

Thurlow, M. L., House, A., Boys, C., Scott, D., & Ysseldyke, J. E. (2000). *State participation and accommodations policies for students with disabilities: 1999 update* (Synthesis Report 33). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L., Hurley, C., Spicuzza, R., & El Sawaf, H. (1996). *A review of the literature on testing accommodations for students with disabilities.* (Minnesota Report 9). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L., Seyfarth, A. L., Scott, D. L., & Ysseldyke, J. E. (1997). *State assessment policies on participation and accommodations for students with disabilities: 1997 update.* (Synthesis Report 29). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L., Ysseldyke, J. E., & Silverstein, B. (1995). Testing accommodations for students with disabilities. *Remedial and Special Education, 16* (5), 260-270.

Tindal, G., & Fuchs, L. (1999). *A summary of research on test changes: An empirical basis for defining accommodations.* Lexington, KY: Mid-South Regional Resource Center.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An experimental study. *Exceptional Children, 64,* 439-450.

Tindal, G., Hollenbeck, K., Heath, B., & Almond, P. (1998). *The effect of using computers as an accommodation in a statewide writing test*. Eugene, OR: University of Oregon.

Turner, G., Tindal, G., Sanford, E., & Chou, F. (1998). Empirical evidence for informing assessment accommodations decisions. Paper presented at the annual CCSSO Large-Scale Assessment Conference, Colorado Springs, CO.

Willingham, W. W., Ragosta, M., Bennett, R. E., Braun, H., Rock, D. A., & Powers, D.E. (1988). *Testing handicapped people*. Needham Heights, MA: Allyn and Bacon.

Ysseldyke, J. E., & Thurlow, M. L. (1994). *Guidelines for inclusion of students with disabilities in large-scale assessments* (Policy Directions 1). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Ysseldyke, J. E., Thurlow, M., Algozzine, B., Shriner, J., & Gilman, C. (1993). *National goals, national standards, national tests: Concerns for all (not virtually all) students with disabilities?* (Synthesis Report 11). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Ysseldyke, J. E., Thurlow, M. L., McGrew, K. S., & Shriner, J. G. (1994). *Recommendations for making decisions about the participation of students with disabilities in statewide assessment programs* (Synthesis Report 15). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Ysseldyke, J. E., Thurlow, M. L., McGrew, K. S., & Vanderwood, M. (1994). *Making decisions about the inclusion of students with disabilities in large-scale assessments* (Synthesis Report 13). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Zlatos, B., (1994). Don't test, don't tell: Is "academic red-shirting" skewing the way we rank our schools? *The American School Board Journal, 181* (11), 24-28.

The College of Education
& Human Development

UNIVERSITY OF MINNESOTA